Weekly Updates (16/11/2023) – Krishna Tarun Saikonda

Driver type Classification - Unsupervised

A total of 1,200 vehicles underwent simulation for 1,000 seconds with a time-step length of 0.1 seconds. This generated an extensive dataset consisting of 400,000 Floating Car Data (FCD) points and 27 distinct attributes. Using this dataset directly for clustering tasks may not be ideal due to the possibility of different vehicle types being grouped together. This could occur because their attributes might match at specific timestamps. For instance, if vehicles of distinct types, such as aggressive and violator, display similar speed, acceleration, or other measurable attributes simultaneously, the clustering algorithm may struggle to differentiate between these types, resulting in inaccurate or misleading groupings.

Data Pre-processing:

To prepare the dataset for a clustering task, a pivotal step was taken to aggregate the information effectively. This process involved grouping the data according to the unique identifiers of each vehicle, namely the ID and the vehicle type. By doing so, the dataset was transformed to capture key features associated with each distinct vehicle. These features encompass various attributes such as the vehicle's position (x, y), angle, speed, acceleration, lane information, and additional characteristics like lane change behavior, lane ID, lane length, vehicle density, and the average speed of nearby vehicles. Aggregating the data in this manner allows for a more meaningful representation of each vehicle, facilitating the identification of patterns and distinctions that might not be apparent when considering individual data points in isolation.

Combining mean, variance, and standard deviation values into a single dataframe is a useful approach to condense and represent statistical features for each vehicle. This consolidation provides a more compact and informative representation of the dataset, potentially enhancing the performance of clustering algorithms. The resulting dataframe likely includes a set of aggregated features for each vehicle, and these features could be crucial for distinguishing patterns among different types of vehicles.

For instance, if the original dataframe contained features like speed, acceleration, and position over time, the aggregated dataframe might now include columns representing the mean, variance, and standard deviation of these attributes for each vehicle. This can help capture the central tendency, spread, and distribution of the vehicle's behavior over the simulated time period.

Such preprocessing steps are common in data analysis and machine learning tasks, as they allow for a more concise and manageable representation of information while retaining essential statistical characteristics. This aggregated dataframe can serve as input for clustering algorithms, enabling them to discern patterns and group vehicles based on their statistical behavior across the recorded attributes.

Unsupervised Learning Models:

1. KMeans Clustering:

- A partition-based clustering algorithm that assigns data points to K clusters based on similarity.

2. Hierarchical Clustering:

- A method that builds a tree of clusters, where each node represents a cluster, and the leaves are individual data points.

Dimensionality Reduction Techniques:

1. PCA (Principal Component Analysis):

- A linear dimensionality reduction technique that aims to capture the most significant variations in the data through orthogonal transformations.

2. t-SNE (t-Distributed Stochastic Neighbor Embedding):

- A non-linear dimensionality reduction technique that emphasizes preserving the local structure of the data, often used for visualization.

3. Autoencoder:

- A neural network-based technique for learning efficient representations of data by encoding and decoding it. In this context, it's used for dimensionality reduction.

Approach:

1. Data Transformation:

- Apply PCA, t-SNE, and Autoencoder to transform the original data into lower-dimensional representations.

2. Clustering Models:

- Use the transformed data from each of the three-dimensional reduction techniques as input to the KMeans and Hierarchical Clustering algorithms.

3. Clustering Analysis:

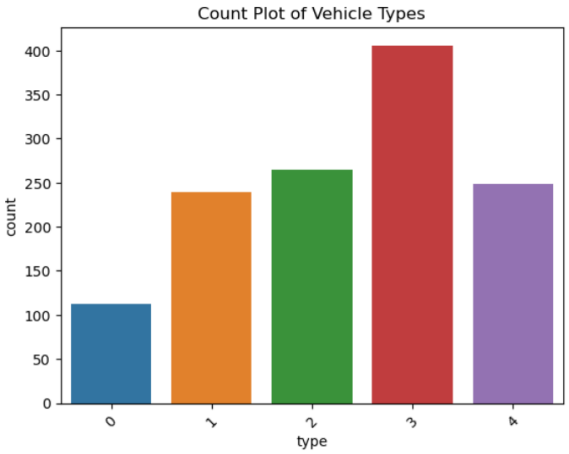
- Evaluate and analyze the results of clustering for each combination of dimensionality reduction technique and clustering algorithm.

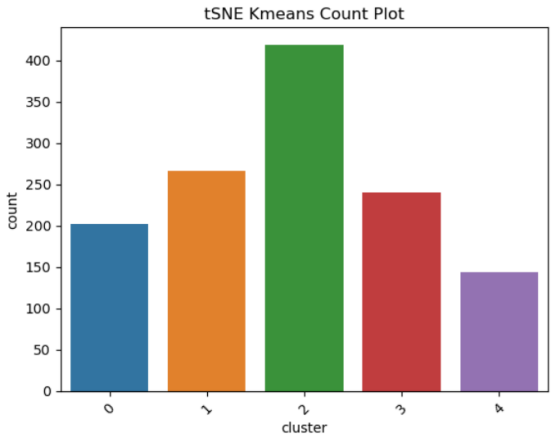
This approach allows you to explore how different dimensionality reduction methods impact the performance of KMeans and Hierarchical Clustering. It's important to assess the quality of the resulting clusters using appropriate metrics (e.g., silhouette score, Davies-Bouldin index) and potentially visualize the clusters to gain insights into the structure of the data.

t-SNE (t-Distributed Stochastic Neighbor Embedding):

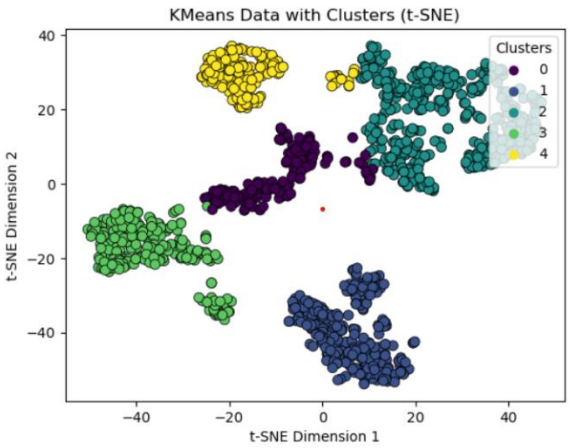
K-Means:

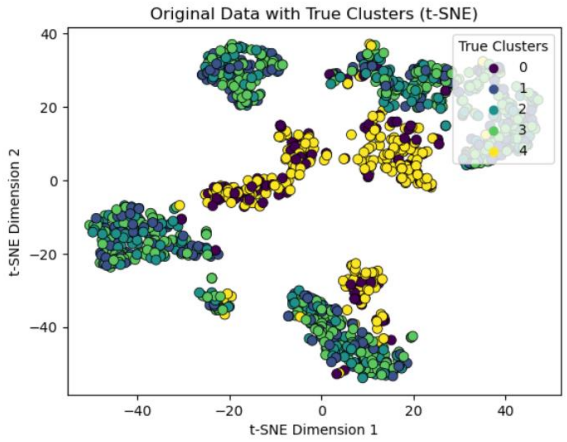
With a Silhouette Score of 0.5707585, the KMeans clustering on t-SNE transformed data indicates a moderately well-defined structure. The distribution of data across clusters reveals varying sizes, with Cluster 2 being the largest (450 points), followed by Clusters 1, 3, 0, and 4. This suggests that certain clusters may exhibit more pronounced patterns in the vehicle data. Further analysis of prominent features within each cluster and potential parameter tuning, such as experimenting with different values of K in KMeans, could provide deeper insights into the underlying characteristics of the data.





The notable difference between the distribution of data points in KMeans clusters and the original data indicates a misalignment. In KMeans clustering, Cluster 3 is the largest, whereas, in the original data, Vehicle Type 3 predominates. This discrepancy suggests that the clustering results may not accurately capture the inherent distribution of vehicle types. Potential causes include the choice of features, the number of clusters (K), or sensitivity to outliers. Further exploration, such as feature engineering or trying alternative clustering methods, may be needed to align the clustering results more closely with the natural grouping of the data.

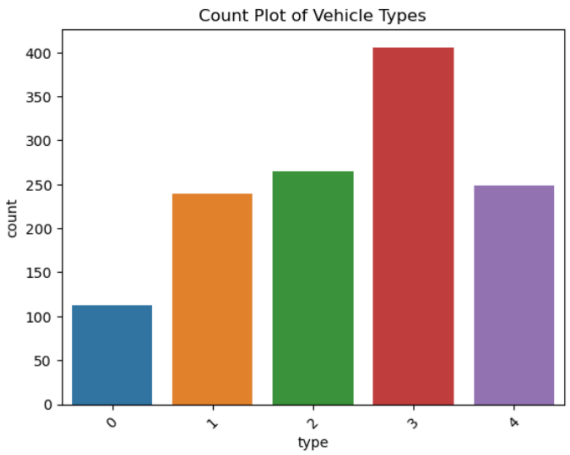


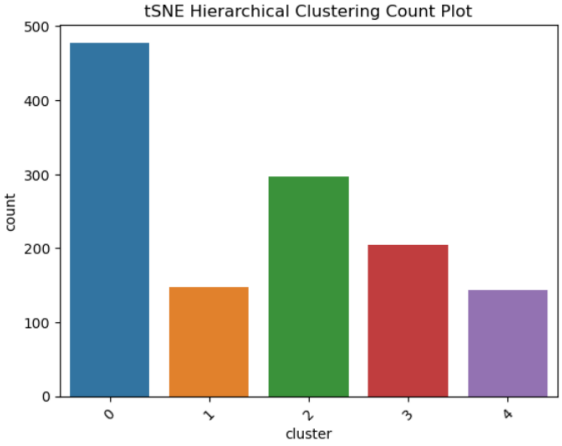


The t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization further supports the existence of well-defined clusters, particularly for clusters 1, 2, and 3, which seem to be more tightly grouped and better separated from each other compared to the PCA visualization.

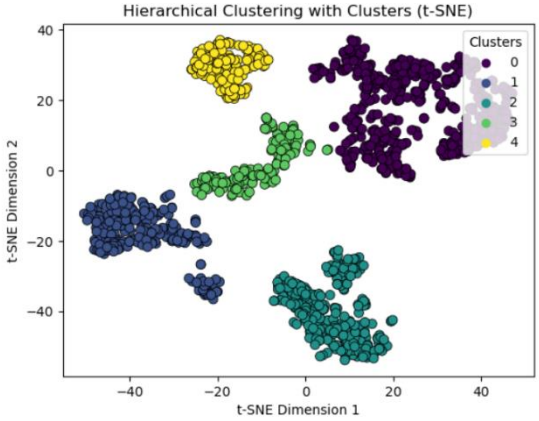
Hierarchical Clustering:

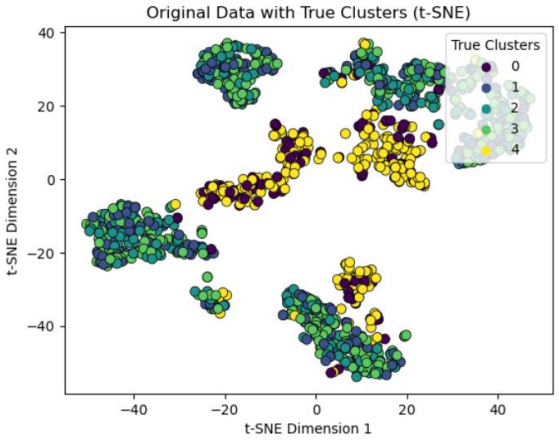
The obtained Silhouette Score of 0.57686466 from Hierarchical Clustering on t-SNE transformed data indicates a reasonably well-defined clustering structure. This score, falling above 0.5, suggests a meaningful separation between clusters. The distribution of data points across clusters shows varying sizes, with Cluster 0 being the largest (477 points), followed by Clusters 2, 1, 3, and 4. This suggests distinct patterns in the vehicle data that are captured by the hierarchical clustering method. The hierarchical approach provides a hierarchical tree structure, and the specific clusters identified may offer insights into the relationships and similarities between different groups of vehicles.





The notable difference in the distribution of data points across clusters between Hierarchical Clustering and the original data implies a discrepancy in the grouping of vehicle types. In Hierarchical Clustering, Cluster 0 is the largest, followed by Clusters 2, 1, 3, and 4. However, when considering the distribution of vehicle types in the original data, Vehicle Type 3 is the most prevalent, followed by types 2, 4, 1, and 0. This misalignment suggests that the hierarchical clustering results may not accurately reflect the natural grouping of vehicle types present in the data.





The t-SNE visualization affirms the presence of distinct clusters, notably in Clusters 1, 2, and 3, which exhibit tighter grouping and clearer separation. This visual alignment corresponds with the results from Hierarchical Clustering, where Cluster 0 is the largest, followed by Clusters 2, 1, 3, and 4. The effectiveness of t-SNE in revealing well-defined clusters reinforces the success of the hierarchical clustering method in capturing the inherent structure of the data.

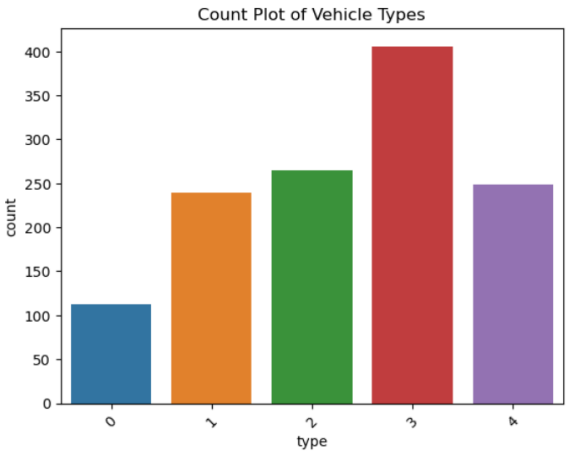
AutoEncoders:

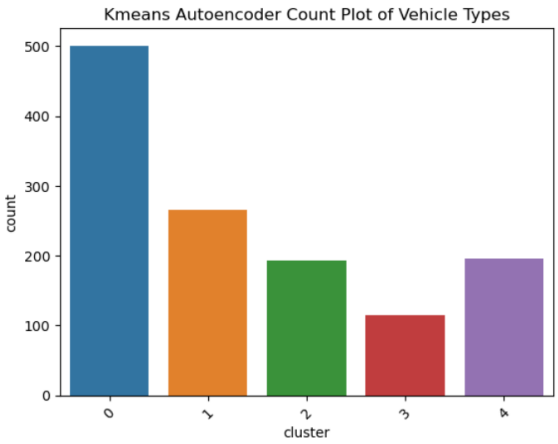
The autoencoder architecture comprises three layers: an input layer, an encoder layer, and a decoder layer. The input layer is defined with a shape corresponding to the number of features in the dataset (38). The encoder layer, implemented as a dense layer with a rectified linear unit (ReLU) activation function, reduces the dimensionality of the input to 2. This encoder layer extracts meaningful representations from the input data. The decoder layer is another dense layer with a sigmoid activation function, aimed at reconstructing the original input dimensionality. The autoencoder model is created using the Keras `Model` class, incorporating the input and output layers. The model is then

compiled using mean squared error (mse) as the loss function and the Adam optimizer. After compilation, the model is trained for 10 epochs with a batch size of 64, using 20% of the data for validation. The final output is the learned encoded representations (`encoded_data`) of the input data.

K-Means:

The obtained Silhouette Score of 0.5637805 from KMeans clustering on the autoencoded data suggests a reasonably well-defined clustering structure. This score, falling above 0.5, indicates a meaningful separation between clusters. The distribution of data points across clusters shows varying sizes, with Cluster 3 being the largest (635 points), followed by Clusters 2, 1, 0, and 4. This clustering outcome implies that the reduced-dimensional representations obtained from the autoencoder have preserved distinctive patterns or structures in the data. The varying sizes of clusters also suggest different levels of prevalence or distinctiveness in the underlying patterns.



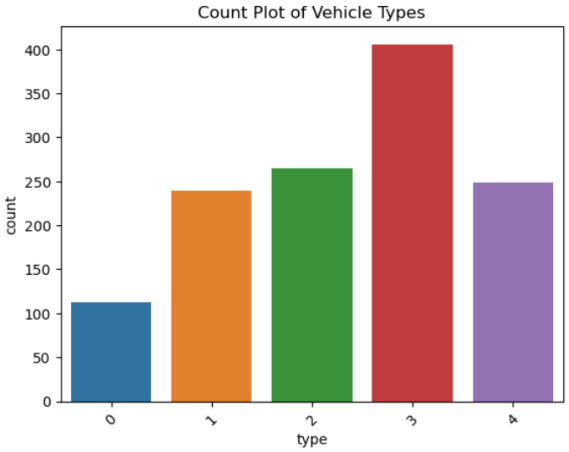


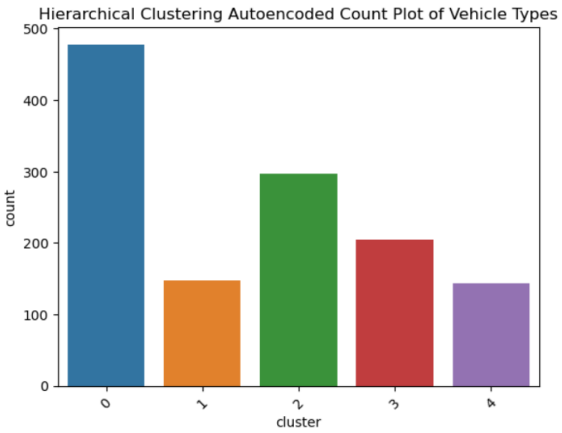
The substantial difference between the distribution of data points across clusters obtained from KMeans clustering on the autoencoded data and the distribution of vehicle types in the original data suggests a misalignment in the clustering results with the inherent grouping of vehicle types. In the KMeans clustering results, Cluster 3 is the largest, followed by Clusters 2, 1, 0, and 4. Conversely, when examining the distribution of vehicle types in the original data, Vehicle Type 3 is the most prevalent, followed by types 2, 4, 1, and 0.

This discrepancy may arise from several factors, including the non-linear transformations introduced by the autoencoder, the choice of the number of clusters (K) in KMeans, or the sensitivity of KMeans to the density and distribution of data points in the reduced-dimensional space. The autoencoder focuses on capturing essential features, and the clustering results might not necessarily align perfectly with the original vehicle types.

Hierarchical Clustering:

The obtained Silhouette Score of 0.25643113 from Hierarchical Clustering on the autoencoded data suggests a moderate, but not strongly defined, clustering structure. This score falls below the typical threshold for well-defined clusters (e.g., 0.5), indicating that the separation between clusters may not be as distinct. The distribution of data points across clusters shows varying sizes, with Cluster 0 being the largest (477 points), followed by Clusters 2, 1, 3, and 4. The hierarchical clustering results imply that the autoencoded data may not exhibit clear-cut hierarchical relationships or separations between clusters.





The noticeable difference between the distribution of data points across clusters obtained from Hierarchical Clustering on the autoencoded data and the distribution of vehicle types in the original data suggests a misalignment in the clustering results with the inherent grouping of vehicle types.

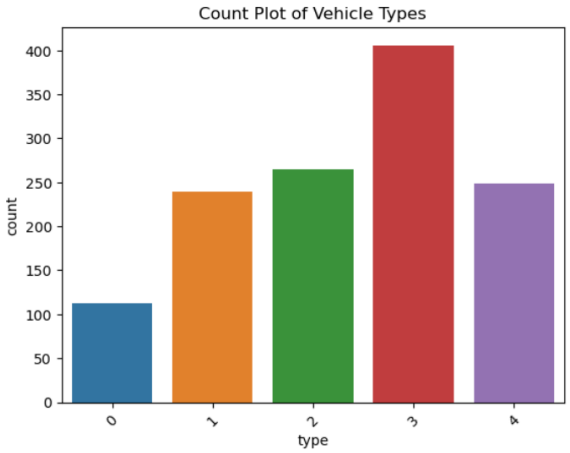
In the hierarchical clustering results, Cluster 0 is the largest, followed by Clusters 2, 1, 3, and 4. However, when considering the distribution of vehicle types in the original data, Vehicle Type 3 is the most prevalent, followed by types 2, 4, 1, and 0.

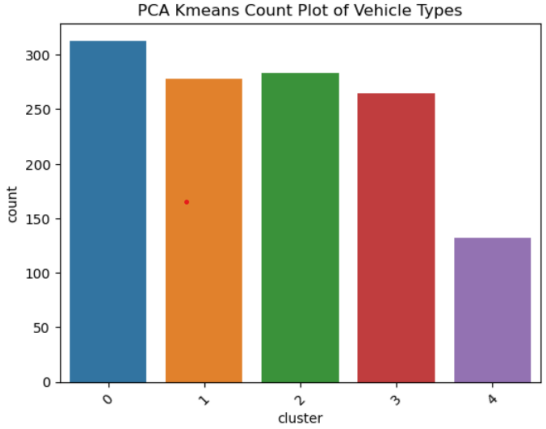
This discrepancy may arise from various factors, including the non-linear transformations introduced by the autoencoder, the choice of linkage criteria or distance metric in hierarchical clustering, or the sensitivity of hierarchical clustering to the density and distribution of data points in the reduced-dimensional space.

PCA (Principal Component Analysis):

K-Means:

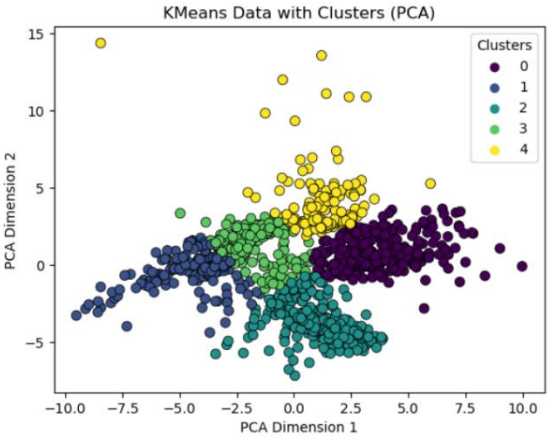
The obtained Silhouette Score of 0.482 from KMeans clustering on the PCA-transformed data suggests a moderate but reasonably well-defined clustering structure. This score, falling between -1 and 1, indicates a meaningful separation between clusters, with a higher score suggesting more distinct clusters. The distribution of data points across clusters shows varying sizes, with Cluster 2 being the largest (450 points), followed by Clusters 3, 0, 1, and 4. This clustering outcome implies that the reduced-dimensional representations obtained from PCA have preserved some distinctive patterns or structures in the data. The varying sizes of clusters suggest different levels of prevalence or distinctiveness in the underlying patterns.

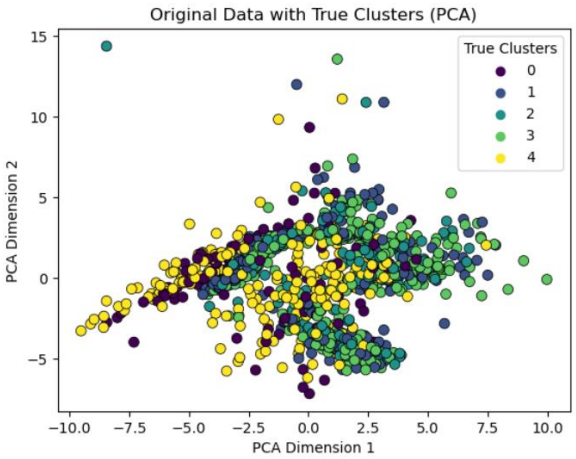




The substantial difference between the distribution of data points across clusters obtained from KMeans clustering on the PCA-transformed data and the distribution of vehicle types in the original data suggests a misalignment in the clustering results with the inherent grouping of vehicle types.

In the KMeans clustering results on PCA data, Cluster 2 is the largest, followed by Clusters 3, 0, 1, and 4. However, when examining the distribution of vehicle types in the original data, Vehicle Type 3 is the most prevalent, followed by types 2, 4, 1, and 0.

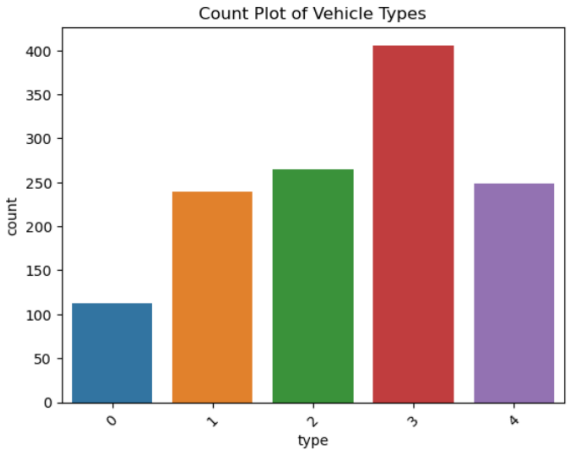


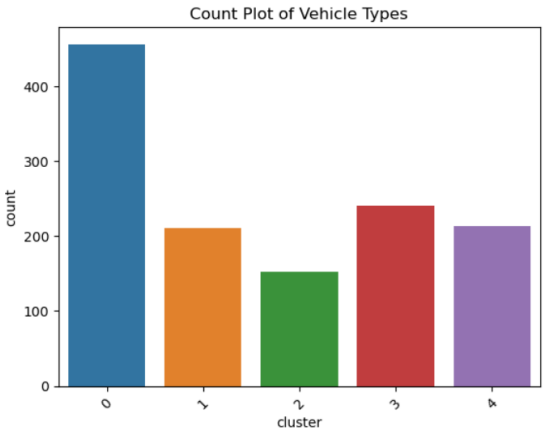


This disparity may arise from the linear nature of PCA, which may not capture the non-linear relationships present in the original data. KMeans clustering on PCA data operates in a reduced-dimensional space, potentially leading to clusters that are more influenced by linear combinations of features rather than the inherent non-linear relationships present in the original data.

Hierarchical Clustering:

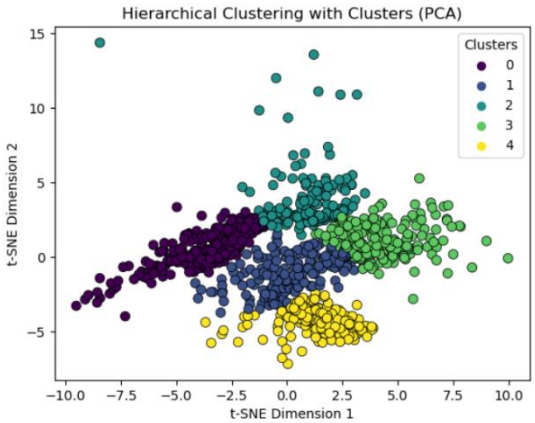
The achieved Silhouette Score of 0.44 from Hierarchical Clustering on the PCA-transformed data suggests a moderate clustering structure. While not as high as some other clustering methods, this score indicates a meaningful separation between clusters. The distribution of data points across clusters exhibits varying sizes, with Cluster 0 being the largest (446 points), followed by Clusters 2, 4, 1, and 3. This clustering outcome suggests that the reduced-dimensional representations obtained from PCA have captured distinct patterns or structures in the data, even though the separation might not be as pronounced as in other clustering methods.

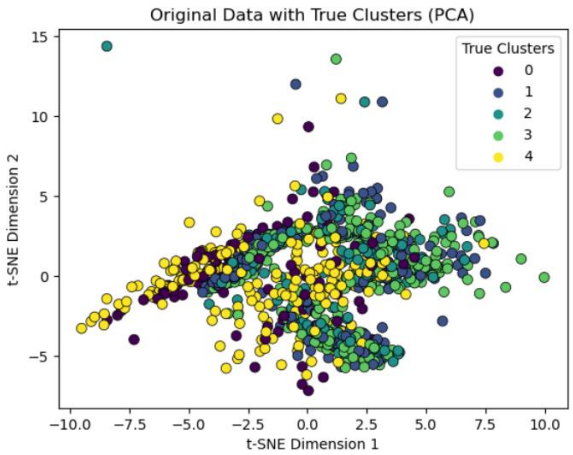




The significant difference between the distribution of data points across clusters obtained from Hierarchical Clustering on the PCA-transformed data and the distribution of vehicle types in the original data suggests a misalignment in the clustering results with the inherent grouping of vehicle types.

In the hierarchical clustering results, Cluster 0 is the largest, followed by Clusters 2, 4, 1, and 3. However, when examining the distribution of vehicle types in the original data, Vehicle Type 3 is the most prevalent, followed by types 2, 4, 1, and 0.





This discrepancy may stem from various factors, including the linear nature of PCA, which may not fully capture the non-linear relationships present in the original data. Hierarchical clustering on PCA data operates in a reduced-dimensional space, potentially leading to clusters that are more influenced by linear combinations of features rather than the intrinsic non-linear relationships present in the original data.